

# Long-read RNA Sequencing Improves the Annotation of the Equine Transcriptome

Peng S.<sup>1</sup>, Dahlgren AR.<sup>1</sup>, Hales EN.<sup>1</sup>, Barber A.M.<sup>2</sup>, Kalbfleisch T.<sup>3</sup>, Petersen JL.<sup>2</sup>, Bellone RR.<sup>1,4</sup>, Mackowski M.<sup>5,6</sup>, Cappelli K.<sup>7</sup>, Capomaccio S.<sup>7</sup>, Coleman SJ.<sup>8</sup>, Distl O.<sup>9</sup>, Giulotto E.<sup>10</sup>, Waud B.<sup>11</sup>, Hamilton N.A.<sup>12</sup>, Leeb T.<sup>13,14</sup>, Lindgren G.<sup>15,16</sup>, Lyons LA.<sup>17</sup>, McCue M.<sup>18</sup>, MacLeod JN.<sup>3</sup>, Metzger J.<sup>9,19</sup>, Mickelson JR.<sup>20</sup>, Murphy BA.<sup>21</sup>, Orlando L.<sup>22,23</sup>, Penedo C.<sup>4</sup>, Raudsepp T.<sup>24</sup>, Strand E.<sup>25</sup>, Tozaki T.<sup>26</sup>, Trachsel DS.<sup>27,28</sup>, Velie BD.<sup>12</sup>, Wade CM.<sup>12</sup>, Cieslak J.<sup>5</sup>, Finno CJ.<sup>\*1</sup>

<sup>1</sup> Department of Population Health and Reproduction, School of Veterinary Medicine, University of California-Davis, Davis, CA, USA

<sup>2</sup> Department of Animal Science, University of Nebraska - Lincoln, Lincoln, NE, USA

<sup>3</sup> Gluck Equine Research Center, Department of Veterinary Science, University of Kentucky, Lexington, USA

<sup>4</sup> Veterinary Genetics Laboratory, School of Veterinary Medicine, University of California-Davis, Davis, CA, USA

<sup>5</sup> Department of Genetics and Animal Breeding, Poznan University of Life Sciences, Poznan, Poland

<sup>6</sup> Horse Genetic Markers Laboratory, Poznan University of Life Sciences, Poznan, Poland

<sup>7</sup> Sport Horse Research Centre, Department of Veterinary Medicine, University of Perugia, Perugia, Italy

<sup>8</sup> Department of Animal Sciences, Colorado State University, Fort Collins, USA

<sup>9</sup> Institute of Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Hannover, Germany

<sup>10</sup> Department of Biology and Biotechnology "Lazzaro Spallanzani," University of Pavia, Pavia, Italy

<sup>11</sup> Sydney School of Veterinary Science, Faculty of Science, University of Sydney, Camperdown, Australia

<sup>12</sup> School of Life and Environmental Science, Faculty of Science, University of Sydney, Camperdown, Australia

<sup>13</sup> Institute of Genetics, Vetsuisse Faculty, University of Bern, Bern, Switzerland

<sup>14</sup> DermFocus, University of Bern, Bern, Switzerland

<sup>15</sup> Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

<sup>16</sup> Center for Animal Breeding and Genetics, Department of Biosystems, KU Leuven, 3001 Leuven, Belgium

<sup>17</sup> Department of Veterinary Medicine and Surgery, College of Veterinary Medicine, University of Missouri - Columbia, Columbia, USA

<sup>18</sup> Veterinary Population Medicine Department, College of Veterinary Medicine, University of Minnesota, Saint Paul, Minnesota, USA

<sup>19</sup> Veterinary Functional Genomics Group, RG Development & Disease, Max Planck Institute for Molecular Genetics, Berlin, Germany

<sup>20</sup> Veterinary and Biomedical Sciences Department, College of Veterinary Medicine, University of Minnesota, Saint Paul, Minnesota, USA

<sup>21</sup> School of Agriculture and Food Science, University College Dublin, Belfield, Dublin, Ireland

<sup>22</sup> Laboratoire d'Anthropobiologie Moléculaire et d'Imagerie de Synthèse, Université de Toulouse, Université Paul Sabatier, Toulouse, France

<sup>23</sup> Lundbeck Foundation GeoGenetics Center, University of Copenhagen, Copenhagen, Denmark

<sup>24</sup> Department of Veterinary Integrative Biosciences, College of Veterinary Medicine and Biomedical Research, Texas A&M University, College Station, USA

<sup>25</sup> Department of Companion Animal Clinical Sciences, Norwegian University of Life Sciences, Oslo, Norway

<sup>26</sup> Genetic Analysis Department, Laboratory of Racing Chemistry, Utsunomiya, Tochigi, Japan

<sup>27</sup> Department of Veterinary and Animal Science, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark

<sup>28</sup> Department of Veterinary Clinical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Taastrup, Denmark

\*Corresponding author: Dr. Carrie J. Finno ([cjfinno@gmail.com](mailto:cjfinno@gmail.com))

## 1 Abstract

A high-quality reference genome assembly, a biobank of diverse equine tissues from the Functional Annotation of the Animal Genome (FAANG) initiative, and incorporation of long-read sequencing technologies, have enabled efforts to build a comprehensive and tissue-specific equine transcriptome. The equine FAANG transcriptome reported here provides up to 45% improvement in transcriptome completeness across tissue types when compared to either RefSeq or Ensembl transcriptomes. This transcriptome also provides major improvements in the identification of alternatively spliced isoforms, novel noncoding genes, and 3' transcription termination site (TTS) annotations. The equine FAANG transcriptome will empower future functional studies of important equine traits while providing future opportunities to identify allele-specific expression and differentially expressed genes across tissues.

## 2 Introduction

Equine genome assemblies<sup>1,2</sup> have provided vital resources for equine genetics research allowing for tools to be built. However, it is evident that detailed annotation of the genome is necessary for further investigation of both simple and complex traits in horses. Current equine genome assemblies have annotations provided by both Ensembl<sup>3,4</sup> and NCBI<sup>5,6</sup> gene annotation pipelines. These annotations relied primarily on limited available RNA-seq data, cross-species

alignments, and computational predictions. The equine RefSeq annotation release 103 from NCBI contains 33,146 genes and pseudogenes, of which 21,129 are protein coding, and 77,102 transcripts, including 60,887 mRNAs and 16,215 non-coding RNAs<sup>6</sup>. This presents a total isoform-to-gene ratio of 2.3, or 2.8 if only coding genes are considered. Similarly, the equine Ensembl annotation release 105.3 contains 20,955 protein coding genes and 9,014 non-coding genes, with 59,087 transcripts, resulting in a transcript-to-gene ratio of 2.0<sup>4</sup>. For comparison, the most recent GENCODE human gene annotation release (release 39, GRCh38.p13) includes 61,533 genes, of which 19,982 are protein coding, with an average isoform-to-gene ratio of 3.9, or 4.3 when only considering protein coding genes<sup>7</sup>. This presents the question of potentially missing alternatively spliced isoforms in the current equine gene annotation. The human ENCODE projects determined that genes tend to express many isoforms simultaneously but different dominant isoforms exist in different cell lines<sup>8</sup>. The tissue-specific nature of isoform expression underscores the urgent need for a transcriptome with more complete isoform annotation.

Additionally, noncoding RNAs play an important role in many biological pathways<sup>9–12</sup>. With the rising popularity of noncoding RNA therapeutics<sup>13</sup>, a comprehensive noncoding RNA annotation for the horse genome will certainly be an asset to the equine research community. The Ensembl annotation for EquCab3 includes 9,014 noncoding genes while RefSeq annotation contains 8,893 noncoding genes. In comparison, GENCODE human gene annotation release 39 includes 26,378 noncoding genes. A particular challenge with annotating noncoding RNAs comes from the fact that noncoding RNAs are usually less evolutionarily conserved<sup>14</sup> and present at very low levels<sup>15,16</sup>. Therefore, without deep sequencing of diverse tissue types, noncoding RNAs typically remain unannotated. Since the current equine gene annotation relies heavily on cross-species conservation, and a limited number of RNA-seq data that are publicly available, it is expected that a large number of noncoding RNAs are currently unannotated.

To address these challenges, the equine Functional Annotation of Animal Genome (FAANG) project has collected over 80 tissue types and body fluids from 4 adult Thoroughbred horses (two females and two males)<sup>17,18</sup>. These horses underwent thorough clinical examinations and were selected as healthy references. The FAANG biobank has produced a diverse dataset describing various aspects of the equine gene regulation<sup>19–21</sup>. Here, we report our efforts to build a comprehensive transcriptome for the horse genome using long-read sequence technologies across an array of diverse tissues.

### 3 Results

#### 3.1 Transcript Annotation

Full-length non-redundant transcripts were categorized based on their annotated splice junctions compared to reference Ensembl transcripts<sup>4,22</sup> and the genomic overlap between the two, following the schematics introduced by Tardaguila *et al*<sup>23</sup>. Overall, isoforms with novel splice sites (novel not in catalog, NNC) account for over 40% of all Iso-seq transcripts (**Table 1**), highlighting the insufficiency of short-read data for splice junction discoveries. The majority of novel genes identified in the Iso-seq transcriptome have only one isoform per gene, which are predominantly mono-exonic (**Fig 1A-D**). Compared to the Ensembl annotation, the Iso-seq transcriptome contains fewer short transcripts (<0.5 Kb) but substantially more long transcripts (>1.5 Kb) (**Fig 1E**).

#### 3.2 5' Completeness

Since standard Iso-seq libraries do not capture 5' caps of transcripts<sup>24</sup>, we used an aggressive collapsing approach where all transcripts with identical 3' ends and only differing at 5' ends are merged into a single transcript. While this approach may hinder the discovery of alternative transcription start sites (TSS), it was necessary to prevent 5' degraded transcripts from being

falsely identified as unique isoforms. To assess the completeness of 5' ends of annotated transcripts, short-read RNA-seq and ATAC-seq data from the same tissues were used to compare coverage near annotated TSS. Overall, 98.4% transcripts have higher RNA-seq coverage in the 100bp window downstream of the Iso-seq annotated TSS than upstream (**Fig 2A**). Transcripts with a  $\log_2$  ratio of greater than 1 were designated as 5' complete. A majority of transcripts across all structural categories were determined to have complete 5' ends, with novel genes (genic, intergenic, and antisense) having a greater percentage of 5' incomplete transcripts (**Fig 2B**). Additionally, ATAC-seq of the same tissues also show substantial enrichment at annotated TSS (**Fig 2C**).

### 3.3 3' Completeness

To capture polyadenylated transcripts with complete 3' ends, poly-T oligonucleotides were used during library construction of Iso-seq. However, internal stretches of adenines could also bind to poly-T oligonucleotides, a phenomenon known as intra-priming, which results in truncated transcripts<sup>23</sup>. To detect potential intra-primed transcripts, the percentage of adenines in a 20 bp window immediately downstream of the annotated transcription termination site (TTS) was calculated for every Iso-seq transcript. Transcripts with 80% or more adenines (i.e., allowing for 4 mismatches with poly-T oligonucleotides) in a 20 bp window downstream of annotated TTS were designated as potential intra-priming candidates. Multi-exonic transcripts across all structural categories had approximately 25% adenines on average, with fewer than 5% transcripts having over 80% adenines, suggesting a high level of 3' completeness (**Fig 3A**). Over 30% of mono-exonic isoforms with novel exons (NIC) are flagged as potentially intra-primed (**Fig 3B**). Many of these transcripts retain a partial intron and may be intron-retaining isoforms undergoing nonsense-mediated decay (NMD). A comparison between Iso-seq and Ensembl annotation showed significant improvement of TTS, as well as minor improvements of TSS annotation for full-splice matched (FSM) transcripts (**Fig 3C**).

### 3.4 Protein Coding and Noncoding Transcripts

Open reading frames (ORFs) were predicted using GeneMarkS-T (GMST) algorithm<sup>25</sup> by SQANTI3<sup>23</sup> to identify protein-coding transcripts in the FAANG transcriptome. The vast majority of transcripts belonging to known genes had ORFs (97.6% of FSM, 96.8% of ISM, 92.5% of NIC, and 95.4% of NNC), while a significant proportion of novel genes had transcripts without ORFs (28% of genic, 60% of intergenic, and 67.6% antisense transcripts) (**Fig 4A**). There is also a substantial difference in exon counts among coding and noncoding transcripts, with 44.6% of noncoding transcripts being mono-exonic as compared to 4.6% of coding transcripts. Specifically, coding transcripts with novel junctions (NIC) are 96.8% multi-exonic while noncoding NIC transcripts are 56% multi-exonic. Similarly, coding transcripts that overlap or fall within annotated introns are 53% multi-exonic, while only 7.5% of those without an ORF are multi-exonic (**Fig 4B**).

### 3.5 Splice Junctions

Any junctions not covered by at least one uniquely mapped read from RNA-seq data were removed, along with their associated transcripts. A total of 8,476 transcripts containing 14,738 such junctions were removed at this step. Known junctions on average had 4.8x RNA-seq coverage as compared to novel junctions. This difference primarily came from canonical junctions (GT-AG, GC-AG and AT-AC) (**Fig 5A**). Novel isoforms and transcripts of novel genes also have lower minimum junction coverage as compared to known isoforms (FSM and ISM, Kruskal-Wallis H-test,  $p < 0.0001$ ; post-hoc Dunn's test  $p < 3.5 \times 10^{-68}$ , Bonferroni corrected  $\alpha = 0.003$ ; **Fig 5B**).

About 10% of the splice junctions annotated in Iso-seq transcriptome were novel compared to Ensembl transcriptome (56,503 out of 581,782). These novel splice junctions contributed discoveries of 36,795 novel isoforms (**Fig 5C**). The GT-AG splice site is observed in 99.2% of

splice junctions, with GC-AG and AT-AC sites observed in 0.68% and 0.05% of transcripts, respectively. Non-canonical splice sites were primarily observed at very low frequencies (<3%) (Fig 5C; Table 2).

### 3.6 Sense-Antisense Transcripts

There were a total of 861 novel antisense transcripts identified, with 2,742 isoforms annotated on the opposite strand. Overall, we identified 3,246 transcripts on the plus strand that overlap at least 1 bp with a transcript on the minus strand. Among these sense-antisense pairs of transcripts, 2,249 (69.3%) were coding-to-coding pairs, 954 (29.4%) coding-to-noncoding pairs, and 43 (1.3%) noncoding-to-noncoding pairs.

### 3.7 Tissue-specific Expression

Short-read RNA-seq data from 57 tissues (46 tissues from female animals and 23 tissues from male animals, with 12 tissues from both sexes, **Supplementary Table 1**) were used to quantify the Iso-seq transcriptome. Most known isoforms were ubiquitously expressed in the majority of tissues sequenced, while novel isoforms of known genes and novel intergenic transcripts each showed a bimodal distribution, with many transcripts detected in only a small number of tissues (Fig 6A). We also noted that the majority of multi-isoform genes expressed more than one isoform in any given tissue (Fig 6B) and had different dominant major isoforms (isoform with highest relative expression of a given gene), depending on the tissue type (Fig 6C-D). Similar to humans, major isoforms in horses usually account for 30%-70% of corresponding genes' total expression in any given tissue (Fig 6E).

Notably, our RNA-seq data exhibited prominent clustering between sexes within the CNS tissues (Fig 7A). Since mare and stallion tissues were prepared at two different laboratories, despite using same protocols, it is unclear whether this clustering reflected genuine biological

differences between sexes or was a result of batch effects arising from the RNA isolations.

Nonetheless, we observed expected clustering of the other major tissue types across all samples (**Fig 7B**).

### 3.8 Comparison with Ensembl and RefSeq Transcriptomes

To assess the completeness of the Iso-seq transcriptome, RNA-seq reads were aligned to the reference genome and portions of the reads that overlapped a transcribed region annotated in the Iso-seq, Ensembl, or RefSeq transcriptome were calculated as fraction of reads in transcript (FRiT). Considerable improvement in FRiT was observed in tissues that were directly utilized to generate the Iso-seq transcriptome (**Fig 8A**). Finally, we aligned all available RNA-seq reads directly to each transcriptome and observed 3-23% improvement in numbers of properly paired reads in all tissues but cerebellum vermis, duodenum, fibroblasts, keratinocytes, bone marrow, and epididymis (caput, corpus, and cauda) (**Fig 8 B**).

To provide a comprehensive set of transcripts for the equine genome, we combined the Iso-seq transcriptome with Ensembl and RefSeq transcriptomes into a single annotation, termed the equine FAANG transcriptome. The FAANG transcriptome consists of 153,492 transcripts (of which 128,723 are multi-exonic) from 36,239 genes, with a gene-to-isoform ratio of 4.2. This combined transcriptome contains a total of 26,631 coding genes, with 132,970 coding transcripts. RNA-seq alignments suggested an average 19.5% (8-45%) improvement in completeness compared to Ensembl and RefSeq transcriptomes across all sequenced tissues (**Fig 8C**).

## 4 Discussion

Here we have presented a comprehensive equine transcriptome with tissue-specific expression utilizing the rich tissue repository from the FAANG biobank and advanced long-read sequencing



technology. The FAANG equine transcriptome consists of 36,239 unique genes with 153,492 transcripts, presenting a gene-isoform ratio of 4.2, or 5.0 when only protein-coding genes are considered. This is a substantial improvement as compared to the ratios of 2.0 from the Ensembl equine transcriptome or 2.8 from the RefSeq equine transcriptome. We also demonstrated improved completeness of our transcriptome across over 40 tissues when assessed by companion RNA-seq data, despite having only 9 tissue types included in our Iso-seq experiment.

The previous efforts to annotate the horse genome were limited by the number of tissue types available and sequencing lengths available at that time<sup>26-28</sup>. Specifically, Hestand *et al.* sequenced 43 different equine tissues in one pool on an Illumina HiSeq 2000, both single- and paired-end at 75 bp on 4 lanes each<sup>27</sup>. While that study included more diverse tissue types than the current FAANG transcriptome, the pooling approach employed in that study limited discovery of rare novel and tissue-specific isoforms. The non-stranded protocol employed in that study also rendered it impossible to identify antisense transcripts. Mansour *et al.* compared sequences of 8 tissue samples from 59 individuals using short-read RNA-seq libraries from several studies (80-125bp, single- and paired-end, stranded and unstranded)<sup>28</sup>. Due to the limitation of short-read sequencing technology, in these studies, it was necessary to aggressively filter out mono-exonic transcripts that were not evolutionarily conserved, a common strategy in short-read based transcriptome assemblies<sup>3,6</sup>. This unfortunately would remove many small noncoding RNAs. Based on recent advances in long isoform sequencing, our approach centered around high-quality full-length reads from Iso-seq and used abundant RNA-seq data to refine splice junction, TSS, and TTS annotation. As a result, we were able to identify many novel mono-exonic transcripts as well as improve TTS annotation for many known transcripts.

In examining the transcriptional pattern of the horse genome, we revealed similar complexity in gene transcription to that of the human<sup>7</sup>. Specifically, we observed that genes with multiple isoforms tend to express more than one isoform simultaneously in any given tissue. The major isoforms (the isoform with highest expression) differed by tissue types. This aligns with the current understanding that isoforms, not genes, are directly associated with tissue-specific biofunctions. Most importantly, our data suggest that most known isoforms annotated in the Ensembl equine transcriptome are ubiquitous, while many novel isoforms identified in the Iso-seq transcriptome show tissue-specific expression. The addition of these novel isoforms should aid the equine genetics community in advancing studies of complex traits.

Despite these improvements, the present Iso-seq transcriptome was unable to accurately define TSS due to a lack of 5' captured reads. Furthermore, small RNAs with no polyadenylated tails are missing from the poly-A captured cDNA libraries used for both Iso-seq and RNA-seq. Assays targeting non-polyadenylated RNAs such as small RNA sequencing and capturing 5' capped transcripts like CAGE-seq are necessary to complement this Iso-seq transcriptome to fully capture the transcriptional landscape in the horse genome. Further, while we demonstrated improved completeness of the FAANG equine transcriptome, only 9 tissues were utilized to construct it, and many rare or tissue-specific transcripts are likely to be missing, especially stem-cell-specific or embryonically specific transcripts. Indeed, bone marrow was the only tissue that showed a drastic decrease in mapping rate when compared to RefSeq or Ensembl transcriptomes. These critical developmental time point and stem cell populations are required to further refine the horse transcriptome.

For the purpose of providing a comprehensive transcriptome, we focused on assessing the completeness of the FAANG equine transcriptome and overall complexities of tissue-specific transcription in the horse. However, the long read-length of Iso-seq data also provides unique opportunities for phasing the exons and, when coupled with whole genome sequencing and

quantifiable RNA-seq of the same animals, studies of allele-specific expression. This is an area of future research. Additionally, the large repository of RNA-seq data from a diverse set of tissues enables studies of differentially expressed genes across tissues. Owing to the addition of the FAANG transcriptome, future studies can focus on quantifying gene expression across tissues and conditions.

## 5 Methods and Materials

### 5.1 RNA Extraction and Sequencing

Since the generation of the FAANG biobank, researchers were invited to “adopt” tissues of their interests, which meant they would sponsor the sequencing costs for two biological replicates (2 male or 2 female) of the “adopted” tissue. Under this Adopt-A-Tissue model, along with the eight prioritized tissues funded by USDA National Institute of Food and Agriculture and the Grayson Jockey Club Foundation, the equine community collectively generated short-read mRNA-seq data from over forty tissues. All RNA extractions for mRNA-seq were performed at the same two locations (female samples at UC Davis, male samples at University of Nebraska-Lincoln).

Briefly, tissue aliquots were homogenized using Biopulverisor and Genogrinder in TRIzol reagent (ThermoFisher Scientific, Waltham MA). RNA was isolated and purified using RNeasy®Plus Mini/Micro columns (Qiagen, Germantown, MD) or Direct-zol RNA Miniprep Plus (Zymo Research, Irvine, CA). A detailed protocol can be found in **Supplemental Materials I and II**. For the female tissues, cDNA libraries were prepared with Illumina TruSeq Stranded kit and sequenced at the University of Minnesota sequencing core facility on an Illumina HiSeq 2500 using 125 bp paired-end reads. Male samples went through similar library preparation before 150 bp paired-end sequencing at Admera Health (South Plainfield, NJ) on an Illumina NovaSeq.

A total of nine tissues (lamina, liver, left lung, left ventricle of heart, longissimus muscle, skin, parietal cortex, testis, and ovary) from the FAANG biobank<sup>17,18</sup> were selected for Iso-seq to represent a wide range of biological functions and therefore, gene expression. RNA for Iso-seq was extracted separately from the same tissues as mRNA-seq using the same protocol. All tissues for Iso-seq were processed in one batch. One sample per sex per tissue was selected for sequencing based on sample availability and RNA integrity numbers (RINs selected > 7). cDNA libraries were prepared and sequenced at UC Berkely QB3 Genomics core facility. Two libraries were randomly pooled and sequenced on a single SMRT cell on PacBio Sequel II.

## 5.2 Transcriptome Assembly

Pooled subreads were first demultiplexed using lima<sup>29</sup>. Circular consensus reads (ccs) were then constructed from demultiplexed subreads using PacBio ccs program<sup>30</sup>. PolyA tails were trimmed from ccs reads using isoseq<sup>31</sup>. This step also removes concatemers and any reads lacking at least 20 bp of polyA tails. Redundant reads were then clustered based on pair-wise alignment using isoseq<sup>31</sup>. Clustered transcripts were aligned to the reference genome EquCab3<sup>2</sup> using minimap2<sup>32</sup> without reference annotation as guide. Collapsed transcripts were filtered if they were not supported by at least two full length reads. Filtered transcripts from each sample were then merged into a single transcriptome using Cupcake<sup>33</sup> and further filtered to retain only those detected in more than one sample. The merged total transcriptome was again aligned to the reference genome and collapsed to remove redundant transcripts. SQANTI3<sup>23</sup> was then used to classify and annotate the transcriptome. Finally, the total transcriptome was filtered to remove nonsense-mediated decay transcripts, transcripts without short-read coverage support, and transcripts with a splice junction not covered by short-read RNA-seq data to generate the final FAANG equine transcriptome. Data processing, visualization, and statistical analyses were performed using pandas<sup>34</sup>, matplotlib<sup>35</sup>, seaborn<sup>36</sup>, scipy<sup>37</sup>, and scikit-learn<sup>38</sup>.

### 5.3 RNA-seq analysis

Short-read RNA-seq data were trimmed to remove adapters and low quality reads using trim-galore<sup>39</sup> and Cutadapt<sup>40</sup>. Read qualities were inspected using fastQC<sup>41</sup> and multiQC<sup>42</sup>. Trimmed reads were aligned to equCab3.0 using STAR aligner<sup>43</sup> with standard parameters (with --outSAMstrandField intronMotif --outSAMattrIHstart 0). PCR duplicates were marked using sambamba<sup>44</sup>. Mapping rates, qualities, and fragment lengths were assessed with samtools<sup>45</sup> and deeptools<sup>46</sup>. Aligned reads were used to assess completeness of transcriptomes using deeptools<sup>46</sup>. BWA MEM<sup>47</sup> was used to align the RNA-seq reads directly to transcriptomes and samtools<sup>45</sup> was used to calculate the percentages of properly paired reads from the transcriptome alignment. Due to the presence of alternatively spliced isoforms in transcriptomes, multiple-alignment reads were not removed.

### 5.4 ATAC-seq analysis

ATAC-seq data from the 8 tissues (lamina, liver, left lung, left ventricle of heart, longissimus muscle, parietal cortex, testis, and ovary) collected from the same animals were generated and processed according to Peng *et al.*<sup>21</sup> Libraries were sequenced in 50 bp paired-end mode (PE50) on Illumina NovaSeq 6000. Aligned reads were used to quantify normalized read counts in 1Kb up- and down-stream of TSS sites annotated in the equine FAANG transcriptome.

## 6 Data Access

RNA-seq data can be accessed from ENA and SRA under the accession number PRJEB26787. Iso-seq data can be accessed from ENA and SRA under the accession number PRJEB53020. ATAC-seq data can be accessed from ENA and SRA under the accession number PRJEB53037. Total and tissue-specific transcriptomes can be downloaded in FASTA and GTF format from <https://equinegenomics.uky.edu/>

## 7 Acknowledgements

The authors would like to acknowledge the four animals from which the FAANG tissues were collected.

## 8 Funding

This project was supported by the Grayson-Jockey Club Research Foundation, Animal Breeding and Functional Annotation of Genomes (A1201) Grant 2019-67015-29340 from the USDA National Institute of Food and Agriculture and the UC Davis Center for Equine Health with funds provided by the State of California pari-mutuel fund and contributions by private donors. Additional support for C.J.F. was provided by NIH L40 TR001136. Funding was also provided through a Priority Partnership Collaboration Award from the University of Sydney and University of California, Davis.

## Conflict of Interest Statement

None

## 9 Figure and Table Legends

**Figure 1. Summary of the FAANG equine transcriptome.** The overall number of isoforms (A), known vs. novel genes (B), percentages of known and novel transcripts (C) and portion of mono- and multi-exon transcripts by structural categories that were annotated in the FAANG equine transcriptome. (D) Percentages of mono- and multi-exonic transcripts in each structural category; FSM: full-splice match, all exons and splice junctions match a known reference; ISM: incomplete-splice match, like FSM but missing 3' and/or 5' ends; NIC: novel-in-catalog, novel transcripts with known exons and splicing sites; NNC: novel-not-in-catalog, novel transcripts

with novel splicing sites; Intergenic: novel transcripts with no overlapping known genes; Genic: novel transcripts overlapping known introns; Fusion: fusion transcripts; Antisense: novel transcripts on the opposite strand of known transcripts (E) The FAANG transcriptome contained fewer short transcripts (<0.5 Kb) but a greater number of long transcripts (>1.5 Kb) as compared to the Ensembl transcriptome.

**Figure 2. 5' Completeness of the FAANG equine transcriptome.** (A) Log<sub>2</sub> of 100 bp downstream over upstream of TSS RNA-seq coverage. Positive ratios indicate higher coverage downstream of TSS. Dotted line indicates equal coverage up- and down-stream of TSS; solid line indicates 100% higher coverage down-stream TSS than up-stream (B) Percentages of transcripts whose log<sub>2</sub> ratios are greater than 1, denoted as 5' complete (green) (C) ATAC-seq read coverage in 1 kb upstream and 3 kb downstream of annotated TSS. FSM: full-splice match; ISM: incomplete-splice match; NIC: novel-in-catalog; NNC: novel-not-in-catalog; Schematics defined by Tardaguila et al<sup>23</sup>.

**Figure 3. 3' Completeness of the FAANG equine transcriptome.** (A) Percentages of adenines in 20 bp genomic regions immediately downstream of annotated TTS. Boxes indicate interquartile range (IQR) and whiskers indicate 1.5\*IQR (B) Portions of transcripts with more than 80% adenines in 20bp genomic regions immediately downstream of annotated TTS, by structural categories and exon counts. (C) Distance between FAANG annotated TTS/TSS and Ensembl annotated TTS/TSS, negative values indicate shorter 5' or 3' ends. Eg. -1000 indicates that FAANG annotated TSS is 1000 bp downstream of Ensembl annotated TSS (left) or that FAANG annotated TTS is 1000 bp upstream of Ensembl annotated TTS (right).

**Figure 4. Protein coding and non-coding transcripts in the FAANG equine transcriptome.**

(A) Portions of coding vs. noncoding transcripts by structural categories and (B) portions of coding vs. noncoding transcripts by structural categories and exon counts.

**Figure 5. Splice junctions are better defined in the FAANG equine transcriptome.**

(A) RNA-seq coverage measured as  $\log_{10}(\text{ReadCount}+1)$  at known or novel splice junctions, (B) minimum splice junction coverage transcripts by structural categories and (C) splice junction types by structural categories, known non-canonical junctions were observed in FSM, ISM, NIC, and NNC at 2.2%, 0.2%, 2.0%, and 1.2%, respectively; novel non-canonical junctions were only observed in NNC and intergenic isoforms at 2.8% and 2.5%, respectively.

**Figure 6. Short-read RNA-seq data mapped to the FAANG equine transcriptome identifies tissue-specific isoforms.**

(A) Distribution of known vs. novel transcripts detected in different numbers of tissues, (B) number of expressed isoforms across all tissues vs. number of total isoforms per gene; boxes indicate IQR and whiskers indicate  $1.5 \times \text{IQR}$ , (C) number of different major isoforms expressed across tissues vs. number of total isoforms annotated, (D) distribution of genes with different number of major isoforms and their total annotated isoforms and (E) relative expression of major isoforms in each tissue vs. total number of isoforms annotated.

**Figure 7. Sex-specific clustering of gene expression across tissue types.**

(A) t-SNE plot of transcript levels (TPM) across samples and tissues and (B) agglomerative clustering of RNA-seq samples.



# **Figure 8. Comparison of FAANG, RefSeq and Ensembl equine transcriptomes. (A)**

Fractions of reads in transcriptome (FRiT) from RNA-seq reads against FAANG, RefSeq, or Ensembl transcriptomes, (B) distribution of percentages of properly paired reads when RNA-seq data is aligned directly to FAANG, RefSeq, or Ensembl transcriptomes and (C) changes in percentages of properly paired reads aligned to combined FAANG transcriptome when compared to Ensembl or RefSeq transcriptomes, whichever has higher percentage.

## **Table 1. FAANG transcripts breakdown by structural category**

FSM: full-splice match; ISM: incomplete-splice match; NIC: novel-in-catalog; NNC: novel-not-in-catalog

<b>Structural Category</b>	<b>Count</b>	<b>Percentage</b>
Antisense	928	1.64
FSM	12470	22.01
Fusion	684	1.21
Genic	2137	3.77
ISM	4937	8.71
Intergenic	4543	8.02
NIC	6330	11.17
NNIC	24634	43.47

## **Table 2. Portions of different splice junction types by structural categories**

FSM: full-splice match; ISM: incomplete-splice match; NIC: novel-in-catalog; NNC: novel-not-in-catalog

	<b>% Known Canonical</b>	<b>% Known Non-canonical</b>	<b>% Novel Canonical</b>	<b>% Novel Non-canonical</b>
<b>Antisense</b>	0.00	0.00	100.00	0.00
<b>Fusion</b>	80.57	0.00	19.43	0.00
<b>Genic</b>	0.00	0.00	100.00	0.00
<b>Intergenic</b>	0.00	0.00	99.97	0.03
<b>NNC</b>	87.35	0.01	12.61	0.03

<b>NIC</b>	96.81	0.02	3.17	0.00
<b>ISM</b>	100.00	0.00	0.00	0.00
<b>FSM</b>	99.98	0.02	0.00	0.00

## 10 References

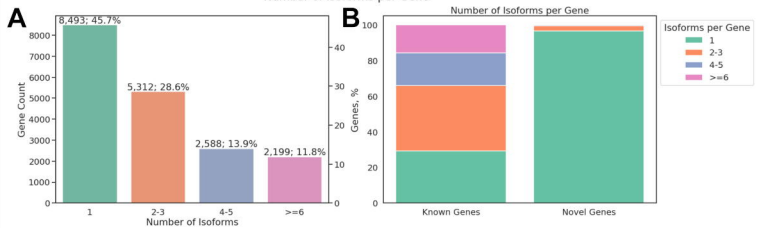
1. Wade, C. M. *et al.* Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science* **326**, 865–867 (2009).
2. Kalbfleisch, T. S. *et al.* Improved reference genome for the domestic horse increases assembly contiguity and composition. *Commun Biol* **1**, 197 (2018).
3. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database (Oxford)* **2016**, (2016).
4. Equus caballus Ensembl Annotation Release 105.3. *Ensembl*  
[https://uswest.ensembl.org/Equus\\_caballus/Info/Annotation](https://uswest.ensembl.org/Equus_caballus/Info/Annotation).
5. O’Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745 (2016).
6. Equus caballus RefSeq Annotation Release 103. *RefSeq*  
[https://www.ncbi.nlm.nih.gov/genome/annotation\\_euk/Equus\\_caballus/103/](https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Equus_caballus/103/).
7. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research* **47**, D766–D773 (2019).
8. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489**, 101–108 (2012).
9. Meller, V. H., Joshi, S. S. & Deshpande, N. Modulation of Chromatin by Noncoding RNA. *Annu. Rev. Genet.* **49**, 673–695 (2015).
10. Shen, Y. *et al.* Insights into Enhancer RNAs: Biogenesis and Emerging Role in Brain Diseases. *Neuroscientist* 107385842110468 (2021) doi:10.1177/10738584211046889.
11. Moazzendizaji, S. *et al.* microRNAs: Small molecules with a large impact on colorectal cancer. *Biotechnology and Applied Biochemistry* bab.2255 (2021) doi:10.1002/bab.2255.

12. Wen, Z.-J. *et al.* Emerging roles of circRNAs in the pathological process of myocardial infarction. *Molecular Therapy - Nucleic Acids* S2162253121002456 (2021)  
doi:10.1016/j.omtn.2021.10.002.
13. Winkle, M., El-Daly, S. M., Fabbri, M. & Calin, G. A. Noncoding RNA therapeutics — challenges and potential solutions. *Nat Rev Drug Discov* **20**, 629–651 (2021).
14. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res.* 22(9):1775–89 (2012).
15. Menet, J. S., Rodriguez, J., Abruzzi, K. C. & Rosbash, M. Nascent-Seq reveals novel features of mouse circadian transcriptional regulation. *eLife* **1**, e00011 (2012).
16. Ramsköld, D., Wang, E. T., Burge, C. B. & Sandberg, R. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLoS Comput Biol* **5**, e1000598 (2009).
17. Burns, E. N. *et al.* Generation of an equine biobank to be used for Functional Annotation of Animal Genomes project. *Anim Genet* **49**, 564–570 (2018).
18. Donnelly, C. G. *et al.* Generation of a Biobank From Two Adult Thoroughbred Stallions for the Functional Annotation of Animal Genomes Initiative. *Front. Genet.* **12**, 650305 (2021).
19. Kingsley, N. B. *et al.* Functionally Annotating Regulatory Elements in the Equine Genome Using Histone Mark ChIP-Seq. *Genes* **11**, 3 (2019).
20. Kingsley, N. B. *et al.* “Adopt-a-Tissue” Initiative Advances Efforts to Identify Tissue-Specific Histone Marks in the Mare. *Front. Genet.* **12**, 649959 (2021).
21. Peng, S., Bellone, R., Petersen, J. L., Kalbfleisch, T. S. & Finno, C. J. Successful ATAC-Seq From Snap-Frozen Equine Tissues. *Front. Genet.* **12**, 641788 (2021).
22. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res* **49**, D884–D891 (2021).

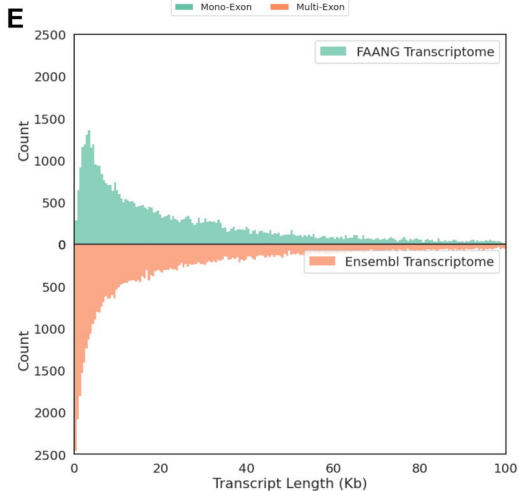
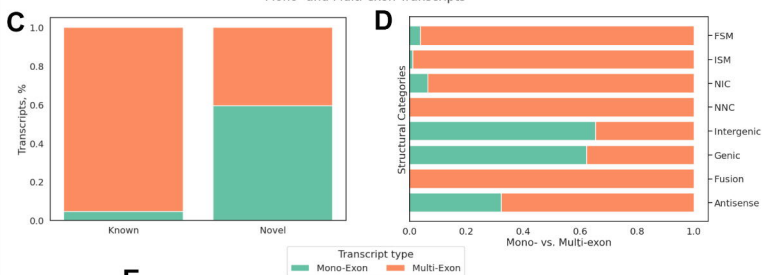
23. Tardaguila, M. *et al.* SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res* (2018) doi:10.1101/gr.222976.117.
24. Gonzalez-Garay, M. L. Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). in *Transcriptomics and Gene Regulation* (ed. Wu, J.) vol. 9 141–160 (Springer Netherlands, 2016).
25. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **43**, e78–e78 (2015).
26. Coleman, S. J. *et al.* Structural annotation of equine protein-coding genes determined by mRNA sequencing: Structural annotation of equine protein-coding genes. *Animal Genetics* **41**, 121–130 (2010).
27. Hestand, M. S. *et al.* Annotation of the Protein Coding Regions of the Equine Genome. *PLoS ONE* **10**, e0124375 (2015).
28. Mansour, T. A. *et al.* Tissue resolved, gene structure refined equine transcriptome. *BMC Genomics* **18**, 103 (2017).
29. *Lima*. (Pacific BioSciences).
30. *ccs*.
31. *isoseq3*.
32. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
33. Tseng, E. *cDNA\_Cupcake*.
34. Reback, J. *et al.* *pandas-dev/pandas: Pandas 1.1.3*. (Zenodo, 2020).  
doi:10.5281/ZENODO.3509134.
35. Caswell, T. A. *et al.* *matplotlib/matplotlib v3.1.3*. (Zenodo, 2020).  
doi:10.5281/ZENODO.3633844.
36. Waskom, M. seaborn: statistical data visualization. *JOSS* **6**, 3021 (2021).

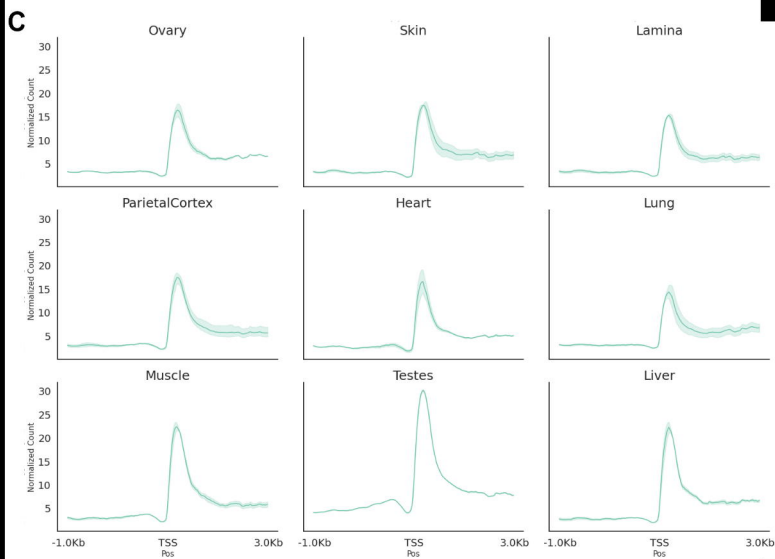
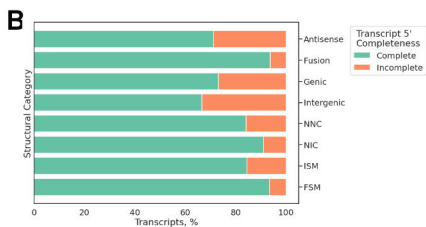
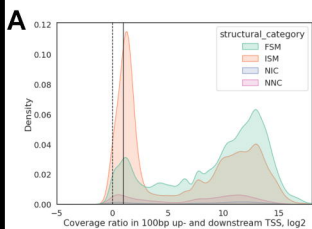
37. SciPy 1.0 Contributors *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, 261–272 (2020).
38. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
39. Krueger, F. *Trim Galore!* (2019).
40. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 10 (2011).
41. Andrews, S. *FastQC: a quality control tool for high throughput sequence data.* (2010).
42. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
43. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
44. Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J. & Prins, P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
45. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
46. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160–W165 (2016).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

Number of Isoforms per Gene

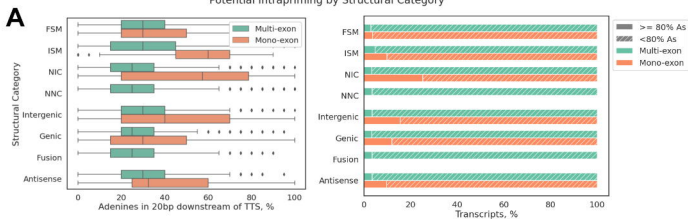


Mono- and Multi-exon Transcripts

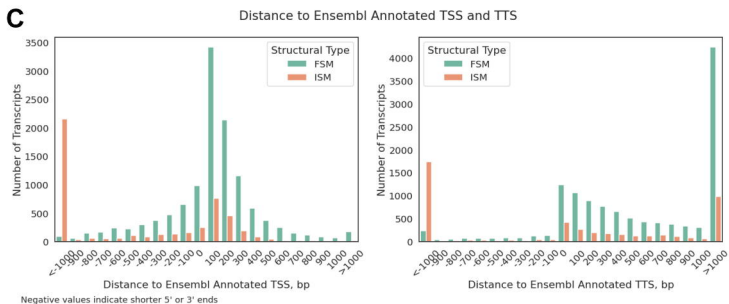




Potential Intrapriming by Structural Category



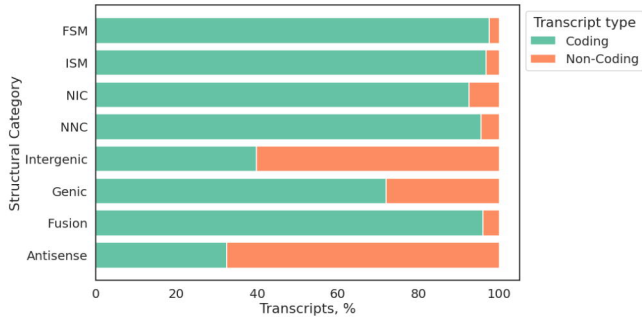
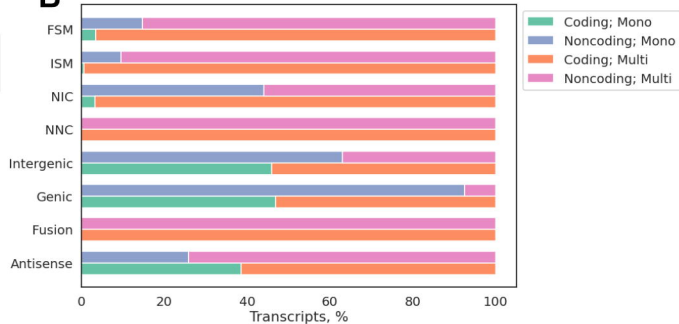
Distance to Ensembl Annotated TSS and TTS

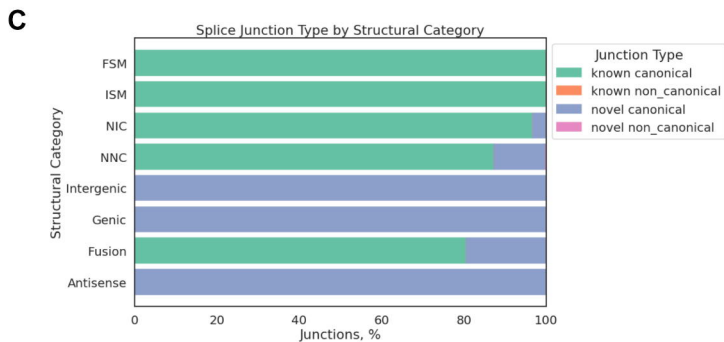
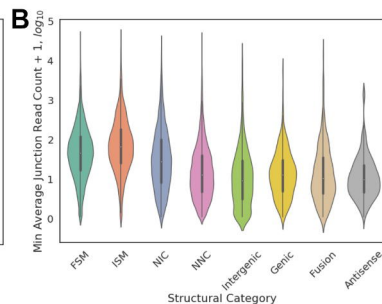
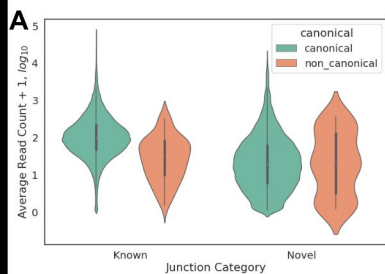


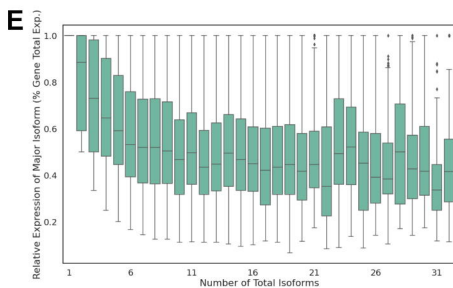
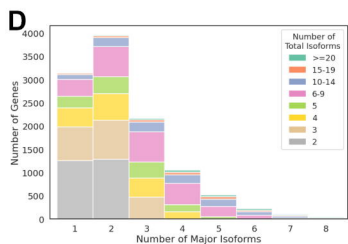
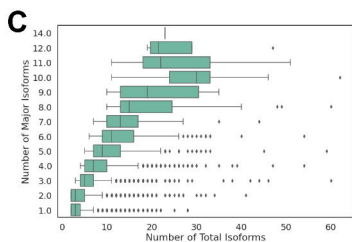
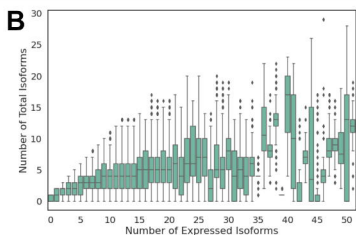
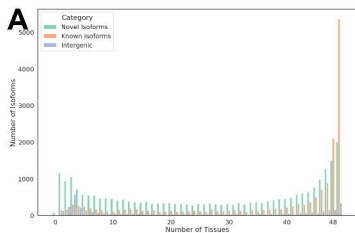


**A**

## Coding vs Non-coding by Structural Category

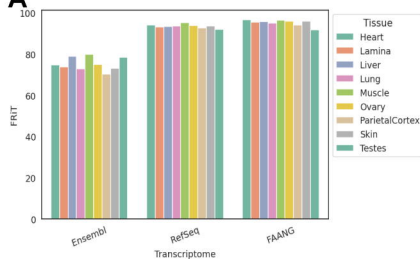
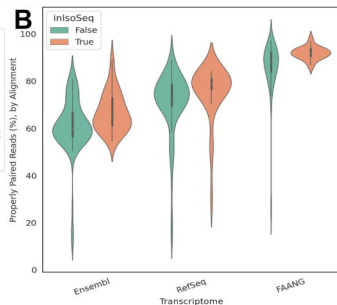
**B**





A t-SNE plot showing single-cell transcriptomic data. The x-axis is labeled 'tSNE-1' and the y-axis is labeled 'tSNE-2'. The plot displays a large cluster of points, with two sub-clusters circled and labeled 'CNS, F' (female) and 'CNS, M' (male). The points are colored by tissue type: Other (blue), Digestive (pink), CNS (green), Liver (yellow), and Lymphatic (orange). The points are also colored by sex: F (female, teal) and M (male, orange). Marginal density plots are shown along the top and right axes, indicating the distribution of points across the t-SNE space.

[illegible]

**A****B****C**